

Bloch spin waves and emergent structure in protein folding with HIV envelope glycoprotein as an example

Jin Dai,^{1,*} Antti J. Niemi,^{2,3,1,†} Jianfeng He,^{1,‡} Adam Sieradzan,^{4,§} and Nevena Ilieva^{5,¶}

¹*School of Physics, Beijing Institute of Technology, Beijing 100081, P.R. China*

²*Department of Physics and Astronomy, Uppsala University, P.O. Box 803, S-75108, Uppsala, Sweden*

³*Laboratoire de Mathématiques et Physique Théorique CNRS UMR 6083,*

Fédération Denis Poisson, Université de Tours, Parc de Grandmont, F37200, Tours, France

⁴*Faculty of Chemistry, University of Gdansk, Wita Stwosza 63, 80-308 Gdańsk, Poland*

⁵*Institute of Information and Communication Technologies,
25A, Acad. G. Bonchev Str., Sofia 1113, Bulgaria*

We inquire how structure emerges during the process of protein folding. For this we scrutinise collective many-atom motions during all-atom molecular dynamics simulations. We introduce, develop and employ various topological techniques, in combination with analytic tools that we deduce from the concept of integrable models and structure of discrete nonlinear Schrödinger equation. The example we consider is an α -helical subunit of the HIV envelope glycoprotein gp41. The helical structure is stable when the subunit is part of the biological oligomer. But in isolation the helix becomes unstable, and the monomer starts deforming. We follow the process computationally. We interpret the evolving structure both in terms of a backbone based Heisenberg spin chain and in terms of a side chain based XY spin chain. We find that in both cases the formation of protein super-secondary structure is akin the formation of a topological Bloch domain wall along a spin chain. During the process we identify three individual Bloch walls and we show that each of them can be modelled with a very high precision in terms of a soliton solution to a discrete nonlinear Schrödinger equation.

PACS numbers: 05.45.Yv, 89.75.Fb, 87.15.hm

INTRODUCTION

A domain wall is a prototype collective excitation in a physical system, and it is also the paradigm example of a topological soliton [1]. A domain wall can appear whenever there is a global symmetry that becomes spontaneously broken. It constitutes the boundary that separates two neighbouring domains, in which the order parameter that detects the symmetry breaking has different values.

In the case of a one dimensional Heisenberg spin chain the order parameter is a three component unit length vector. When one of the three vector components vanishes identically, the Heisenberg spin chain reduces to the XY spin chain [2, 3]. A domain wall along the spin chain is a localised excitation that interpolates between two different, ordered spin states in which the order parameter has different constant values. Two major types of domain walls are commonly identified along the Heisenberg chain [2, 3]. These are called the Bloch wall and the Néel wall, respectively. In the case of a Bloch wall, the Heisenberg spin variable rotates through the plane of the wall and in the case of a Néel wall the rotation takes place within the plane of the wall itself. Domain walls that are mixtures of these two, can also occur along the Heisenberg spin chain, while along the XY spin chain, only domain walls of the Bloch type can be present.

In this article we demonstrate that the formation of super-secondary structures, during folding of a protein [4], can be understood in terms of a Bloch domain wall

that forms along a Heisenberg spin chain, or along a closely related XY spin chain. We propose that the spin chain interpretation of a protein backbone provides a systematic framework for understanding and describing the process of protein folding. For this we employ all-atom force fields [5, 6] to scrutinise protein folding dynamics at the level of individual atoms and their oscillations. We analyse the folding pathway using a combination of topological techniques and global analytic tools. We isolate the collective oscillations which are pertinent for the folding process, from the noisy background of thermal and random individual atom fluctuations. In particular, we illustrate how the individual atom motions become organised and combined into a coherent structural excitation which we identify as the Bloch wall.

As a concrete example we consider an α -helical subunit of the HIV envelope glycoprotein gp41 [7], with Protein Data Bank [8] (PDB) code 1AIK. There are six α -helical subunits in the biological assembly, shown in Figure 1. We consider in isolation the subunit, for which the first amino acid has number 628 in the PDB file. In isolation, the subunit is unstable and starts folding.

The transmembrane glycoprotein 41 is itself a subunit of the retrovirus envelope protein complex. In the case of the HIV, its structure has been studied extensively. It is presumed to have substantial biological relevance to the initial viral infection. Accordingly, the gp41 protein is a popular target for the development of an anti-viral immune response, to prevent and cure HIV infection. However, medical applications are beyond the direct scope of

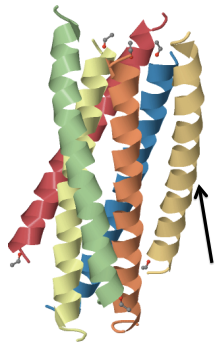


FIG. 1: The biological assembly of 1AIK is an oligomer with six α -helical structures. The subunit that starts with amino acid number 628 in the PDB file is identified by the arrow.

the present study. Here we shall solely address and identify the physical mechanism, why and how an individual, initially α -helical subunit of 1AIK becomes unstable in isolation, and starts folding.

For our all-atom molecular dynamics simulations, we utilise the GROMACS 4.6.3 package [9]. We analyse the results using a variety of topological techniques and analytical tools. Our approach derives from the mathematical structure of Heisenberg and XY spin chains, in combination with properties of a discrete nonlinear Schrödinger (DNLS) Hamiltonian [2, 3]. In particular, the DNLS equation that describes the local extrema of the Hamiltonian, enables us to analytically identify the profile of the domain wall, and to interpret it in terms of DNLS soliton [10].

Here we present results from the detailed investigation of a particular example. However, we expect our observations and conclusions to be generic. Indeed, the present results are fully in line with the previous findings [11–13] obtained by using the coarse grained UNRES energy function [14–16] in the case of protein A. The similarity of results that are obtained by analysing the protein folding process using different tools, built and based on phenomena with very different characteristic time and length scales, demonstrates that we have correctly identified the relevant collective motions that command the folding process.

METHODS

We have performed *in silico* experiments to fold one C-chain subunit of the core structure of gp41 [7]. The structure comes from the HIV envelope glycoprotein with PDB code 1AIK. The amino acid sequence is

$$\begin{array}{l} \text{W M E W D R E I N N Y T S L I H S} \\ \text{L I E E S Q N Q Q E K N E Q E L L} \end{array} \quad (1)$$

These amino acids are assigned the numbers 628-661 in the PDB entry of 1AIK.

All-atom simulations

We have used the molecular dynamics package GROMACS 4.6.3 [9]. We have analysed in detail a number of 80 ns long trajectories, with the crystallographic PDB conformation as the initial condition. We have chosen the length of the trajectories by inspecting, when major structural deformations take place. We have employed three different force fields, to eliminate force-field based artifacts. These are the united-atom force field GROMOS53a6, and the all-atom force fields CHARMM27 and OPLS/AA.

The 1AIK subchain that we have investigated in detail, consists of 34 amino-acid residues, with PDB numbers 628-661. There are 16200 atoms in the entire system that we have simulated, including the solvent. The simulation box has dimensions $47 \times 47 \times 74 \text{ \AA}^3$. This ensures that there is a 2 nm minimal distance between the protein atoms and the box walls, with periodic boundary conditions.

We have described the solvent using the SPC water model [17]. We have neutralised the system at a salt concentration of 0.15 mol/l. We have used steepest-descent for initial energy minimisation. The system was warmed up to 290 K by a simulated annealing in a 100 ps position-restraint simulation. We have chosen this relatively low temperature value for a better control of random thermal noise but without forgoing the underlying physical phenomena. For temperature control we have employed the Berendsen-thermostat with a time constant 0.1 ps, and for pressure coupling — the Berendsen-barostat with a pressure set to 1 bar and a time constant 0.5 ps. Constraints on all bonds were imposed with the LINCS algorithm [18]. We have used the particle mesh Ewald (PME) method [19] to compute the long-range electrostatic interactions, with van der Waals and Coulomb cutoff radii of 0.9 nm. For the 80 ns production run with a time step of 2 fs, that we analyse here in detail, we have changed the thermostat to v-rescale and the barostat to Parrinello-Rahman, keeping the initial time constants, to ensure the generation of a proper canonical ensemble [9]. We have recorded the coordinates every 20 ps, which gives rise to 4000 frames that form the basis for our analysis.

Protein geometry

We have introduced, employed and developed a number of topological tools and analytic techniques to analyse and interpret the results of our GROMACS simulations.

Discrete Frenet equation

We monitor the evolution of the protein geometry using Frenet frames which are based on the backbone C α atoms [20]. The framing depends *only* on the C α atom coordinates \mathbf{r}_i , where $i = 0, \dots, N$ labels the residues and $N = 33$ in the case of 1AIK. At a given \mathbf{r}_i the frame consists of the unit backbone tangent (\mathbf{t}_i), binormal (\mathbf{b}_i) and normal (\mathbf{n}_i) vectors, defined as follows,

$$\mathbf{t}_i = \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{|\mathbf{r}_{i+1} - \mathbf{r}_i|} \quad (2)$$

$$\mathbf{b}_i = \frac{\mathbf{t}_{i-1} \times \mathbf{t}_i}{|\mathbf{t}_{i-1} \times \mathbf{t}_i|} \quad (3)$$

$$\mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i \quad (4)$$

Our aim is to identify and isolate the collective multi-atom motions that drive the protein folding process, from the background of the various random fluctuations. We expect that such coherent motions and oscillations have characteristic time scales, which are much longer than the period of an individual atom covalent bond oscillation. In average, over the relevant time scales, the distance between two consecutive C α atoms can then be taken to be nearly constant, and equal to

$$|\mathbf{r}_{i+1} - \mathbf{r}_i| \approx 3.8 \text{ \AA} \quad (5)$$

Thus, at relevant time scales, the backbone dynamics can be described entirely in terms of the virtual backbone bond and torsion angles κ_i and τ_i , as the *complete structural order parameters* [21, 24]. These angles are defined as follows,

$$\kappa_{i+1,i} \equiv \kappa_i = \arccos(\mathbf{t}_{i+1} \cdot \mathbf{t}_i) \quad (6)$$

$$\tau_{i+1,i} \equiv \tau_i = \omega \arccos(\mathbf{b}_{i+1} \cdot \mathbf{b}_i) \quad (7)$$

where

$$\omega = \text{sign}[(\mathbf{b}_{i-1} \times \mathbf{b}_i) \cdot \mathbf{t}_i] \quad (8)$$

Conversely, the frame vectors (2)-(4) can be expressed in terms of these two order parameters iteratively, using the *discrete Frenet equation* [20]

$$\begin{pmatrix} \mathbf{n}_{i+1} \\ \mathbf{b}_{i+1} \\ \mathbf{t}_{i+1} \end{pmatrix} = \begin{pmatrix} \cos \kappa \cos \tau & \cos \kappa \sin \tau & -\sin \kappa \\ -\sin \tau & \cos \kappa & 0 \\ \sin \kappa \cos \tau & \sin \kappa \sin \tau & \cos \kappa \end{pmatrix}_{i+1,i} \begin{pmatrix} \mathbf{n}_i \\ \mathbf{b}_i \\ \mathbf{t}_i \end{pmatrix} \quad (9)$$

and the C α backbone is calculated from

$$\mathbf{r}_k = \sum_{i=0}^{k-1} |\mathbf{r}_{i+1} - \mathbf{r}_i| \cdot \mathbf{t}_i \quad (10)$$

Unlike the tangent vector \mathbf{t}_i , the normal and binormal vectors \mathbf{n}_i and \mathbf{b}_i do not appear in equation (10). Thus,

if we rotate these two vectors simultaneously around the vector \mathbf{t}_i , the C α geometry remains intact and only the way how it is framed changes. In particular, rotation by π constitutes the discrete \mathbb{Z}_2 gauge transformation

$$\begin{aligned} \kappa_i &\rightarrow \kappa_i - \pi \\ \tau_k &\rightarrow -\tau_k \end{aligned} \quad \text{for all } k \geq i \quad (11)$$

This transformation has been previously used extensively, to analyse protein loop structure [10–13, 20, 21, 24–27]. It will also be used in the sequel.

Heisenberg spin variables

According to (10) the entire C α backbone geometry is determined by the tangent vectors \mathbf{t}_i . Thus, following [28–30] we may visualise the backbone geometry in terms of these vectors: We take the base of \mathbf{t}_i to be at the location \mathbf{r}_i of the i^{th} C α atom. We identify the tip of \mathbf{t}_i as a point on the surface of a unit two-sphere \mathbb{S}_i^2 that is centered at the point \mathbf{r}_i . We orient the coordinate system on the sphere so that the north-pole coincides with the tip of \mathbf{t}_i . Thus, the north-pole is always in the direction of the next C α , which is at the site \mathbf{r}_{i+1} .

We proceed to characterise the direction of the next tangent vector \mathbf{t}_{i+1} *i.e.* the direction from \mathbf{r}_{i+1} towards the C α atom at site \mathbf{r}_{i+2} , in terms of the longitude and latitude angles of the i^{th} two-sphere \mathbb{S}_i^2 . For this, we translate the center of \mathbb{S}_i^2 from \mathbf{r}_i towards its north-pole, and all the way to the location \mathbf{r}_{i+1} of the $(i+1)^{\text{th}}$ C α atom, without introducing any rotation of the sphere. We then record the direction of \mathbf{t}_{i+1} as a point on the surface of the translated \mathbb{S}_i^2 . This defines the coordinate values (κ_i, τ_i) , that determine how the backbone chain turns at site \mathbf{r}_{i+1} , to reach the $(i+2)^{\text{th}}$ central C α atom at the point \mathbf{r}_{i+2} : The angle κ_i measures the latitude of \mathbf{t}_{i+1} on the translated two-sphere \mathbb{S}_i^2 , from its north pole. The angle τ_i measures the longitude of \mathbf{t}_{i+1} , starting from the great circle that passes both through the north pole and through the tip of the binormal vector \mathbf{b}_i .

When we repeat the above procedure for all C α atoms, we obtain a (κ, τ) distribution that characterises the overall geometry of a protein backbone. For a visualisation of this distribution we employ the geometry of a stereographically projected two-sphere: We project the (κ, τ) coordinates from the south pole to the tangent plane of the north pole, of the two-sphere. If (x, y) are the coordinates of this tangent plane the projection is defined by

$$x + iy = \tan\left(\frac{\kappa}{2}\right) \cdot e^{-i\tau} \quad (12)$$

When we perform the projection for all C α atoms in all crystallographic protein structures in PDB that have been measured with resolution better than 2.0 Å, we arrive at the statistical angular distribution that we show in

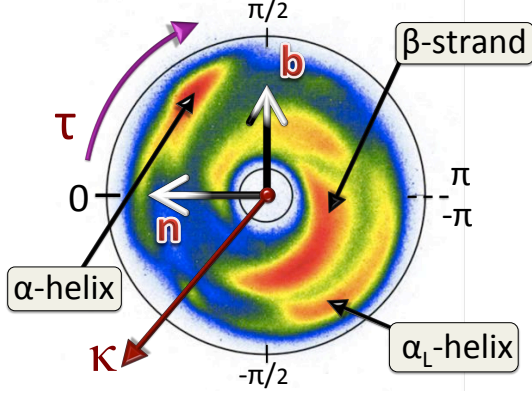


FIG. 2: The distribution of (κ, τ) values in all PDB structures with better than 2.0 Å resolution, on the stereographically projected two-sphere, with a rainbow encoding of the number of entries (red corresponding to the largest number). The locations of the major regular secondary structures are identified.

Figure 2. It is the landscape for the shape of the protein backbones from the crystallographic data in PDB. By the way it is obtained, the crystallographic protein structure should be very close to a stationary minimum of the ensuing Gibbs free energy. Thus the figure 2 should be the collective landscape of stationary, minimum-energy protein structures.

We observe that the PDB data is concentrated in an annulus which is roughly between the circles $\kappa_{in} \approx 1$ and $\kappa_{out} \approx \pi/2$. The exterior of the annulus $\kappa > \kappa_{out}$ is an excluded region, the ensuing conformations are subject to steric clashes. The interior $\kappa < \kappa_{in}$ is sterically allowed but in practice excluded in PDB structures. Note that regular structures such as α -helices and β -strands are distinguished as highly localised regions in the figure 2, with

$$(\kappa, \tau)_\alpha \approx (1.57, 0.87) \sim \left(\frac{\pi}{2}, 1\right)$$

for α -helices and

$$(\kappa, \tau)_\beta \approx (1, -2.9) \sim (1, \pm\pi)$$

for β -strands. Different regions in Figure 2 can be connected by loops, which can be considered as trajectories along the variables (κ, τ) . We have found that loops have the tendency to encircle the inner circle. In figure 3 a) we show, as an example, a generic loop that connects the right-handed α -helical region with the β -stranded region.

To describe a backbone segment analytically, we combine its $C\alpha$'s bond and torsion angles into the three component unit vectors

$$\mathbf{s}_i = \begin{pmatrix} \cos \tau_i \sin \kappa_i \\ \sin \tau_i \sin \kappa_i \\ \cos \kappa_i \end{pmatrix} \quad (13)$$

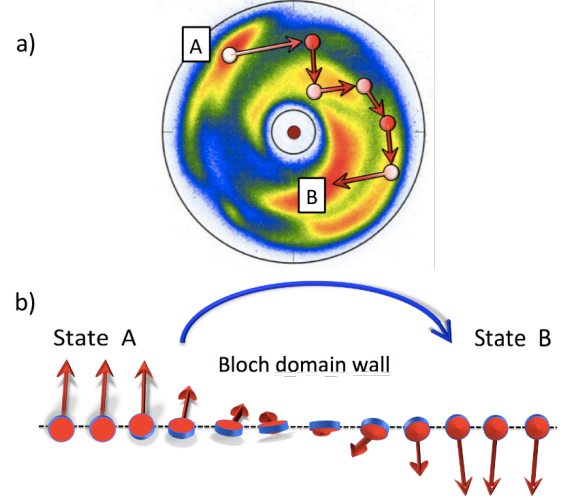


FIG. 3: a) A generic loop is a trajectory on the stereographically projected (κ, τ) sphere, that connects a region corresponding to a regular secondary structure (here A) to another one (here B). b) In terms of the variable (13), a loop becomes a Bloch domain wall that interpolates between ground states A and B, along a Heisenberg spin chain.

We interpret these vectors as the local order parameters along an imaginary *linear* one dimensional Heisenberg spin chain, labeled by the index i . This converts the $C\alpha$ geometry into a configuration along a linear Heisenberg chain in a one-to-one manner: In Figure 3(b) we have sketched how the (generic) trajectory shown in Figure 3(a) appears, figuratively, in terms of such a Heisenberg spin chain configuration.

Since the spin variable (13) takes values only in the annulus $\kappa_{in} < \kappa < \kappa_{out}$ of the two-sphere S^2 , it is apparent that a loop can be *de facto* identified as a domain wall akin the Bloch wall along a Heisenberg chain. The loop then interpolates between the two different regular secondary structures, denoted by state A and state B respectively, as shown in the figures 3.

Residues and spin chains

The amino acid side chains can be similarly interpreted in terms of a one dimensional linear spin chain. In fact, there are several ways to identify the spin chain variable. Here we utilise the directional vector that points from the $C\alpha$ atom at \mathbf{r}_i towards the ensuing $C\beta$ atom, located at \mathbf{r}_i^β . This vector can be introduced for all amino acids except glycine (G); note that there is no glycine in (1).

We start with the unit vector

$$\mathbf{u}_i^\beta = \frac{\mathbf{r}_i^\beta - \mathbf{r}_i}{|\mathbf{r}_i^\beta - \mathbf{r}_i|} \quad (14)$$

We recall the $C\alpha$ based discrete Frenet framing with the

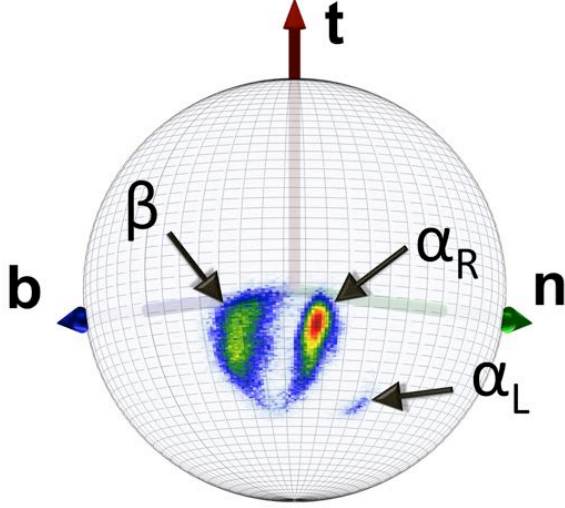


FIG. 4: $C\beta$ distribution in the corresponding $C\alpha$ centered discrete Frenet frames for all structures in PDB with resolution better than 1.0 Å. The regions corresponding to α -helices (α_R), β -strands (β), left-handed α -helices (α_L) are identified, the rest are (mostly) loops.

coordinates (κ_i, τ_i) and represent (14) as three component unit vectors in this coordinate system,

$$\mathbf{u}_i^\beta \rightarrow \hat{\sigma}_i = \begin{pmatrix} \cos \tau_i^\beta \sin \kappa_i^\beta \\ \sin \tau_i^\beta \sin \kappa_i^\beta \\ \cos \kappa_i^\beta \end{pmatrix} \quad (15)$$

Here $(\kappa_i^\beta, \tau_i^\beta)$ are the spherical coordinates of the i^{th} $C\beta$ atom, on the surface of the $C\alpha$ centered two-sphere \mathbb{S}_i^2 . In Figure 4 we show the distribution of the vectors (15) on the surface of the two-sphere, for all those crystallographic PDB structures that have been measured with resolution better than 1.0 Å. Note that the sphere is the same as in figures 2 and 3(a) but now there is no stereographic projection.

We can interpret the distribution in Figure 4 as the $C\beta$ landscape of stationary folded protein structures with minimum Gibbs energy. The highly localized character of the distribution shows that there is a very strong correlation between the $C\alpha$ (backbone) geometry and the $C\beta$ (side chain) geometry. Accordingly, the ground state structures of the corresponding Heisenberg spin chain Hamiltonians must be very similar.

We proceed to introduce a set of $O(2)$ spin variables for the side-chain $C\beta$. For this we define the projection of (14) onto the normal plane at the position of the i^{th} $C\alpha$,

$$\mathbf{u}_i = \frac{\mathbf{u}_i^\beta - (\mathbf{u}_i^\beta \cdot \mathbf{t}_i) \mathbf{t}_i}{|\mathbf{u}_i^\beta - (\mathbf{u}_i^\beta \cdot \mathbf{t}_i) \mathbf{t}_i|}$$

For the next $C\beta$ along the chain, we introduce similarly the vector \mathbf{u}_{i+1} and compute its projection onto the *same*

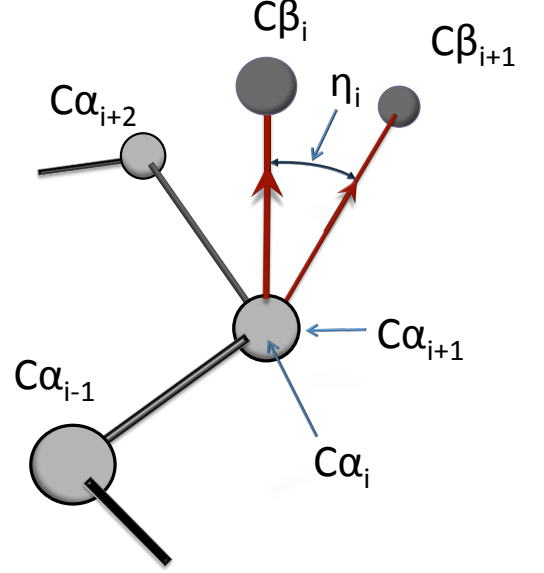


FIG. 5: The angle η_i in (16) is defined as the angle between the projections of the vectors \mathbf{u}_i and \mathbf{u}_{i+1} , connecting the i^{th} $C\alpha$ and $C\beta$, and the $(i+1)^{st}$ $C\alpha$ and $C\beta$ on the normal plane of \mathbf{t}_i . Note that in the figure the i^{th} $C\alpha$ is in front of (on top of) the $(i+1)^{st}$ $C\alpha$.

normal plane — at the position of the i^{th} $C\alpha$,

$$\mathbf{v}_i = \frac{\mathbf{u}_{i+1}^\beta - (\mathbf{u}_{i+1}^\beta \cdot \mathbf{t}_i) \mathbf{t}_i}{|\mathbf{u}_{i+1}^\beta - (\mathbf{u}_{i+1}^\beta \cdot \mathbf{t}_i) \mathbf{t}_i|}$$

We then define the relative angle η_i ,

$$\cos \eta_i = \mathbf{u}_i \cdot \mathbf{v}_i \quad (16)$$

As shown in Figure 5, η_i is the dihedral angle

$$\eta_i := C\beta(i) - C\alpha(i) - C\alpha(i+1) - C\beta(i+1) \quad (17)$$

We note that the construction resembles that of Newman projection in stereochemistry.

In analogy with Figure 3(b) we identify the variable η_i as an order parameter for a linear $O(2)$ XY spin chain,

$$\mathbf{m}_i = \begin{pmatrix} \cos \eta_i \\ \sin \eta_i \end{pmatrix} \quad (18)$$

Like the Heisenberg model, the XY model supports domain walls that interpolate between two configurations where the order parameter has different constant values. The domain wall of the XY model is akin the Bloch domain wall of a Heisenberg model. Figure 4 shows that the ground state structure of the side chain XY model is closely related to that of the backbone Heisenberg model, in the case of crystallographic PDB protein structures.

Folding indices

The formation, evolution and structure of a loop along a folding protein can be monitored in terms of topologically determined folding indices. Here we are interested in two particular examples of folding indices, one that relates to the backbone geometry and another one that relates to the side chain geometry.

In the case of a Heisenberg spin chain, there is a topological index akin a winding number that characterises and classifies its Bloch domain walls. For the $C\alpha$ Bloch wall shown in figure 3 a), b) this topological index counts the net number of times the corresponding trajectory encircles around the annulus in the figure *i.e.* around the north-pole of the two-sphere. We remind that due to steric constraints, the Heisenberg variable (13) takes values in the annulus shown in figures 2, 3 a). We also recall that for the first homotopy class of a circle $\pi_1(\mathbb{S}^1) \simeq \mathbb{Z}$ which justifies the introduction of a topological concept.

Analytically, we may assign to each loop, and more generally to a backbone segment, between residues n_1 and n_2 the following folding index Ind_f [31],

$$Ind_f = [\Gamma] \quad (19)$$

where

$$\Gamma = \frac{1}{\pi} \sum_{i=n_1+2}^{n_2-2} \begin{cases} \tau_i - \tau_{i-1} - 2\pi & \text{if } \tau_i - \tau_{i-1} > \pi \\ \tau_i - \tau_{i-1} + 2\pi & \text{if } \tau_i - \tau_{i-1} < -\pi \\ \tau_i - \tau_{i-1} & \text{otherwise} \end{cases} \quad (20)$$

Here $[x]$ denotes the integer part of x . Note that Γ is the total rotation angle (in radians) that the projections of the $C\alpha$ atoms of the consecutive loop (segment) residues make around the north pole. The n_1 , n_2 label the first and last residue of the loop. Commonly these are the last *resp.* first residues in the preceding and in the following regular secondary structures. The folding index is a positive integer when the rotation is counterclockwise, and a negative integer when the rotation is clockwise. The folding index can be used to classify individual loop structures and backbone segments, even entire protein backbones [31]. Note that the folding index is normalised so that it is equal to twice the number of times the vector in Figure 3(b) rotates around its axis, when the spin structure traverses a domain wall *i.e.* it assigns an even integer to the $\pi_1(\mathbb{S}^1) \simeq \mathbb{Z}$ winding number, in the case of a closed trajectory.

For example, for the trajectory shown in figures 3 the folding index has the value -1. For a loop connecting an α -helix and a β -strand, the folding index is generically an odd integer. For a loop connecting two α -helices, or two β -strands, the folding index is generically an even integer.

In the case of the side chains, we utilise the $C\beta$ angular XY spin variable (16) to define a similar topological

folding index [13]. For this, we first choose a reference residue, *e.g.* the n^{th} residue along the backbone. Starting with this reference residue, we then evaluate the accumulated *total* angle $\hat{\eta}_m$ over a segment with $m - n$ residues,

$$\hat{\eta}_m = \sum_{k=n}^m \eta_k \quad (21)$$

and we define the ensuing index by

$$Ind_m = \left[\frac{\hat{\eta}_m}{\pi} \right] \quad (22)$$

Again, the index acquires its topological justification from the fact that $\pi_1(\mathbb{S}^1) \simeq \mathbb{Z}$. The dihedral η_k , the accumulated total angle (21) and the ensuing index (22) can all be used to study and classify loop structures, protein segments, and entire proteins.

Landau free energy

A generic all-atom molecular dynamics simulation of a folding protein contains a wide range of intermediate conformations. Some of these are essential for the correct folding pathway, while some are merely random transients with no inherent relevance to the folding process *per se*. In order to identify the relevant conformational processes, we need systematic methods that smooth over and weave out the irrelevant random fluctuations. For this we recall the standard Wilsonian universality arguments [22, 23], to deduce the form of the Landau free energy that emerges from the thermodynamical Gibbs free energy. In the present context, the derivation is based on the following two assumptions [24]:

- We assume that the characteristic length scales that are associated with spatial variations and deformations along the protein backbone around its thermal equilibrium configuration, are large in comparison to the covalent bond lengths. This presumes that there are no abrupt edges but only gradual slowly varying bends and twists along the backbone. From Figures 2 and 4 we conclude that the steric constraints between the backbone and the side chain atoms act as powerful inhibitors of sharp, edgy motions.

- We also assume that the individual $C\alpha$ virtual bond length oscillations have characteristic time scales, which are very short in comparison to the time scale which characterise a folding process. The characteristic time scale of a random covalent bond oscillation is around ten femtoseconds while in our simulations we record the individual atomic coordinates every 20 picosecond. We have tested that conformational changes which take place over shorter time scales, do not affect our conclusions. Accordingly, we may adopt (5) as the (time averaged) value for all the nearest neighbour $C\alpha$ - $C\alpha$ distances.

It has been shown that in the case of crystallographic PDB structures, the bond and torsion angles (κ_i, τ_i) form

a complete set of structural order parameters [21]. Accordingly, in the vicinity of a Gibbs free energy minimum we may expand the free energy in terms of these angles. For this we consider the response of the interatomic distances to variations in these angles, with

$$r_{ab} = r_{ab}(\kappa_i, \tau_i)$$

where r_{ab} is the distance between *any* two C α atoms a and b along the backbone.

Suppose that at a given local extremum of the free energy, the C α bond and torsion angles have the equilibrium values

$$(\kappa_i, \tau_i) = (\kappa_{i0}, \tau_{i0})$$

We then consider a non-equilibrium conformation where the (κ_i, τ_i) deviate from these extremum values. We denote the deviations by

$$\begin{aligned} \Delta\kappa_i &= \kappa_i - \kappa_{i0} \\ \Delta\tau_i &= \tau_i - \tau_{i0} \end{aligned} \quad (23)$$

When the deviations are slowly varying in space, *i.e.* (23) are small, we may Taylor expand the Gibbs free energy around the extremum,

$$\begin{aligned} G(r_{\alpha\beta}) &\equiv G[r_{\alpha\beta}(\kappa_i, \tau_i)] = \\ &G(\kappa_{i0}, \tau_{i0}) + \sum_k \left\{ \frac{\partial G}{\partial \kappa_k|_0} \Delta\kappa_k + \frac{\partial G}{\partial \tau_k|_0} \Delta\tau_k \right\} + \\ &+ \sum_{k,l} \left\{ \frac{1}{2} \frac{\partial^2 G}{\partial \kappa_k \partial \kappa_l|_0} \Delta\kappa_k \Delta\kappa_l + \frac{1}{2} \frac{\partial^2 G}{\partial \tau_k \partial \tau_l|_0} \Delta\tau_k \Delta\tau_l \right. \\ &\quad \left. + \frac{\partial^2 G}{\partial \kappa_k \partial \tau_l|_0} \Delta\kappa_k \Delta\tau_l \right\} + \mathcal{O}(\Delta^3) \end{aligned}$$

The first term in the expansion evaluates the free energy at the extremum. Since (κ_{i0}, τ_{i0}) correspond to the extremum, the second term vanishes. Thus we are left with the following expansion of the averaged free energy,

$$\begin{aligned} G(\kappa_i, \tau_i) &= G(\kappa_{i0}, \tau_{i0}) \\ &+ \sum_{k,l} \left\{ \frac{1}{2} \frac{\partial^2 G}{\partial \kappa_k \partial \kappa_l|_0} \Delta\kappa_k \Delta\kappa_l + \frac{1}{2} \frac{\partial^2 G}{\partial \tau_k \partial \tau_l|_0} \Delta\tau_k \Delta\tau_l \right. \\ &\quad \left. + \frac{\partial^2 G}{\partial \kappa_k \partial \tau_l|_0} \Delta\kappa_k \Delta\tau_l \right\} + \dots \end{aligned} \quad (24)$$

When the characteristic length scale of spatial deformations around a minimum energy configuration is large

in comparison to a covalent bond length, we may *re-arrange* the expansion (24) in terms of the differences in the angles as follows: first come local terms, then come terms that connect the nearest neighbours, then come terms that connect the next-to-nearest neighbours, and so forth. When we re-order the expansion (24) in this manner and demand that the free energy is invariant under local rotations in the $(\mathbf{b}_i, \mathbf{n}_i)$ -plane, we conclude [24, 32, 33] that to the leading order the expansion of the Gibbs free energy *necessarily* coincides with the energy of the following discrete nonlinear Schrödinger equation [2, 3, 10]

$$\begin{aligned} F = \sum_{i=1}^N \left\{ \lambda (\kappa_i^2 - m^2)^2 + \frac{q}{2} \kappa_i^2 \tau_i^2 - p \tau_i + \frac{r}{2} \tau_i^2 + \dots \right\} \\ + \sum_{i=1}^{N-1} (\kappa_{i+1} - \kappa_i)^2 + \dots \end{aligned} \quad (25)$$

$$\equiv V_{pot}[\kappa, \tau] + \sum_{i=1}^{N-1} (\kappa_{i+1} - \kappa_i)^2 \quad (26)$$

This functional form of the free energy is simply the most general Landau free energy that one can write down using the available variables (κ_i, τ_i) , in a manner which is consistent with the symmetry principle that a local rotation of the $(\mathbf{n}_i, \mathbf{b}_i)$ frames has no effect on the backbone geometry. The corrections to (25) include next-to-nearest neighbours couplings and so forth, which are higher order terms from the point of view of our systematic expansion.

We note that the expansion (25) has the property that in continuum limit it yields the Coleman-Weinberg derivative (low momentum) expansion [34]

$$F \rightarrow \int_0^L ds \{ V(\phi) + A + |(\partial_s + iA)\phi|^2 + \dots \} \quad (27)$$

where, following [32, 33], we have identified the bond angle with the complex scalar field $\kappa_i \rightarrow \phi(s)$ and the torsion angle with the U(1) gauge field $\tau_i \rightarrow A(s)$, in the continuum limit.

We *emphasize* that the approximation (25) is valid in the limit of slow spatial variations (low momentum). That is, as long as there are no abrupt, sharp edges but only gradual bends and twists along the backbone. In particular, long range interactions are accounted for as long as they do not cause any localised sharp buckles along the backbone, and the angular variations respect the steric constraints.

The Wilsonian universality arguments are sufficient to conclude that in the limit of slowly varying backbone geometry *any* complete all-atom force field can be approximated by the energy function (25). The parameters

λ , q , p , r , and m depend on the atomic level physical properties and the chemical microstructure of the protein and its environment. In principle, these parameters can be computed from this knowledge. But as always in the case of a Landau free energy, it remains a challenge to compute these parameters from the all-atom level.

Spontaneous symmetry breaking and solitons

The free energy (25) relates to the DNLS energy function [2, 3, 10]. The non-linear, quartic bond angle contribution is the familiar double-well potential that gives rise to a spontaneous breakdown of the \mathbb{Z}_2 symmetry

$$\kappa_i \longleftrightarrow -\kappa_i$$

The spontaneous breakdown of this discrete symmetry is pivotal for the emergence of a loop structure, in the case of proteins. It gives rise to a Bloch wave that interpolates between the two ground states $\kappa_i = \pm m$.

More generally, the quartic potential admits a non-symmetric profile of the form

$$U \approx \sum_{C\alpha} \frac{1}{2} k_0 (\kappa - a)^2 (\kappa - b)^2 \quad (28)$$

Here a and b are the positions of the minima of the quartic potential, and k_0 is a force constant. By carefully taking the continuum limit of the $C\alpha$ lattice, *i.e.* the limit where (5) becomes small, and by introducing a mass-scaled variable ξ with m representing the effective mass of a residue, the pertinent DNLS equation becomes

$$m \frac{d^2 \xi}{ds^2} = -k_0 \xi (\xi^2 - c^2) \quad (29)$$

where s is the arc length parameter along the backbone. With $c = (a + b)/2$ and

$$m\xi = \kappa - \frac{1}{2}(a + b) \quad (30)$$

the solution of equation (29) is

$$\xi(s) = c \tanh \left[c \sqrt{\frac{k_0}{2m}} (s - s_0) \right] \quad (31)$$

where s_0 is the position of the inflection point, *a.k.a.* the center of a kink. In terms of the original variables and parameters

$$\kappa(s) = \frac{b e^{c \sqrt{\frac{k_0}{2m}} (s-s_0)} + a e^{-c \sqrt{\frac{k_0}{2m}} (s-s_0)}}{\cosh \left[c \sqrt{\frac{k_0}{2m}} (s - s_0) \right]} \quad (32)$$

This is known as the dark soliton solution of the non-linear Schrödinger equation. It interpolates between the

asymptotic values which correspond to the (local) minima of the potential,

$$\kappa(s) \rightarrow \begin{cases} a & s \rightarrow -\infty \\ b & s \rightarrow +\infty \end{cases} \quad (33)$$

In the case of a protein, the soliton describes the bond angle profile of a super-secondary structure such as (α -helix)-(loop)-(β -strand) shown in figure 3; the parameters have the values $a \approx 1.5$ and $b \approx 1.1$ (radians) for the states A and B , shown in the figure.

In the case of a protein chain, the arc length s becomes replaced by a discrete variable which is equal to the position of the ensuing $C\alpha$ in the sequence. The variables κ_i and τ_i are also mutually interacting, according to (25). The soliton is constructed as the minimum of F in equation (25) [10, 24–26]. It is the solution of a system of $2N - 5$ nonlinear equations in $2N - 5$ unknowns, where N is the number of residues. In order to obtain the solution, we first solve for τ_i in terms of κ_i ,

$$\tau_i[\kappa] = \frac{p}{r + q \kappa_i^2} \equiv \frac{u}{1 + v \kappa_i^2} \quad (34)$$

with $u = p/r$ and $v = q/r$. By inserting equation (34) into equation (25), the torsion angles τ are eliminated and we obtain a system of equations for the bond angles κ ,

$$\kappa_{i+1} = 2\kappa_i - \kappa_{i-1} + \frac{dV_{pot}[\kappa]}{d\kappa_i^2} \kappa_i \quad (i = 1, \dots, N) \quad (35)$$

where $\kappa_0 = \kappa_{N+1} = 0$ and

$$V_{pot}[\kappa] = \frac{p}{r + q \kappa^2} + 2(1 - \lambda m^2) \kappa^2 + \lambda \kappa^4 \quad (36)$$

Here we recognize the discretised structure of equation (29). The difference is in the first term on the right-hand side in equation (36). However, it turns out that in the case of proteins, its effect is not that pronounced as the effect of the other terms; it turns out that the first term is small in value when compared to the other two.

We can construct the profile of the dark soliton solution to equation (35) numerically, by following the iterative procedure introduced in reference [10]; the explicit form of the solution is until now unknown, in terms of elementary functions. However, we obtain an *excellent* approximation [26] by *naively* discretising the continuum dark nonlinear Schrödinger equation soliton (32)

$$\kappa_i = \frac{\mu_1 \exp[\sigma_1(i - s)] + \mu_2 \exp[-\sigma_2(i - s)]}{\exp[\sigma_1(i - s)] + \exp[-\sigma_2(i - s)]} \quad (37)$$

Here $\mu_{1,2} \in [0, \pi] \bmod(2\pi)$ are parameters, which determine the amplitude of the variation of κ and the asymmetry of the inflection regions. The parameters σ_1 and σ_2 are related to the inverse of the range of the inflection region. We remark that in the case of proteins, the

values of $\mu_{1,2}$ are determined entirely by the adjacent helices and strands. Furthermore, far away from the soliton center we have in analogy with (33)

$$\kappa_i \rightarrow \begin{cases} \mu_1 & \text{mod } (2\pi) \quad i > s \\ \mu_2 & \text{mod } (2\pi) \quad i < s \end{cases}$$

The corresponding torsion angles are evaluated in terms of the bond angles using equation (34).

Note that in the case of proteins, the profile of equation (37) becomes monotonically increasing when we add multiples of 2π to the experimental values. Since the values of κ_i 's are defined $\text{mod } (2\pi)$, this does not affect the backbone geometry. The integer number of times the monotonically increasing variable κ_i covers its fundamental domain $[-\pi, \pi)$ counts the number of solitons along the backbone. Recall that negative values of κ_i are related to positive values of κ_i by \mathbb{Z}_2 symmetry (11). Finally, *only* the parameters σ_1 and σ_2 in (37) are intrinsically specific parameters for a given loop. But they specify only the length of the loop, not its shape which is determined *entirely* by the functional form of equation (37) and, as in the case of μ_1 and μ_2 , they are combinations of the parameters in equation (36).

In the expression (34) of the torsion angles τ_i , $i = 1, 2, \dots, N-3$, there are only two independent parameters u and v . Consequently the profile of τ_i is determined entirely by κ_i , and by the structure of the adjacent regular secondary structures.

It has been shown [27] that most crystallographic protein structures in PDB can be described with very high precision in terms of such soliton profiles as their modular building blocks. Moreover, it has been found that in the ensuing soliton profiles, the number of parameters is generically much smaller than the number of residues. Thus, the energy function (25) has a very high predictive power, in describing folded proteins structures in PDB. Its predictions can be subjected to stringent experimental scrutiny, both in the case of static and dynamic proteins.

RESULTS

We now proceed to demonstrate, that all the concepts and structures we have identified are observed during an all-atom simulation of protein folding. We start with individual atom level scrutiny, even though our goal is to identify and model those *collective* conformational deformations that cause a protein to fold. We inquire how does self-organisation, in the case of a protein, relate to universal concepts such as formation of domain walls along spin chains. We study how accurately can the dynamics and structure of the important collective deformations be modelled by soliton profiles such as the one described by the DNLS equation.

We have subjected the single C-chain subunit of the core structure of gp41 [7] with PDB code 1AIK and

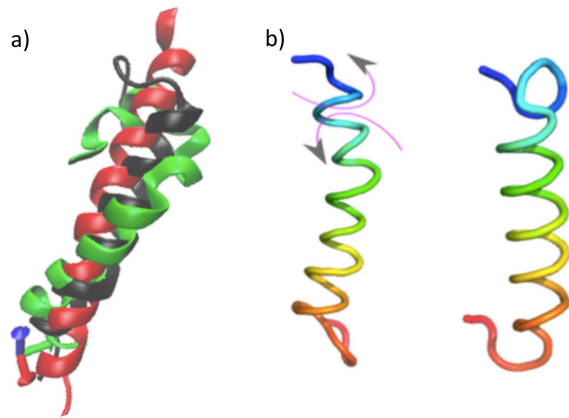


FIG. 6: a) Final conformations after 80 ns MD simulations with: GROMOS53a6 force field (black); CHARMM27 force field (red), and OPLS/AA force field (green). b) During the process that we simulate, we commonly observe that the N-terminal rotates anticlockwise while the rest of the protein rotates clockwise.

residues 628-661, to detailed all-atom simulations. We have used the GROMACS 4.6.3. package [9] with three different force fields — GROMOS53a6, CHARMM27 and OPLS/AA — having thus the protein described by 383, 573 and 609 atoms, respectively. The final production simulation models 80 ns of the protein evolution. We have concluded this to be sufficient, to identify and analyse the important structural deformations that can take place.

Generalities

Backbone

In Figure 6(a) we show the secondary structure of the final conformations that we have obtained using the three force fields, in our 80 ns simulations. We observe a clear deformation, in a segment that consists of the first 10 residues from the N-terminal (upper part in the figure).

In Figure 7 we display the weighted root-mean-square-deviation (RMSD) of the protein backbone in the three force fields,

$$\text{RMSD}(t_1, t_2) = \left[\frac{1}{M} \sum_{i=1}^n m_i \|\mathbf{r}_i(t_1) - \mathbf{r}_i(t_2)\|^2 \right]^{1/2} \quad (38)$$

Here $\mathbf{r}_i(t)$ is the position of the atom i at time t , and

$$M = \sum_{i=1}^n m_i$$

where m_i are the individual atom masses, and M is the total mass of the backbone. The deformation is

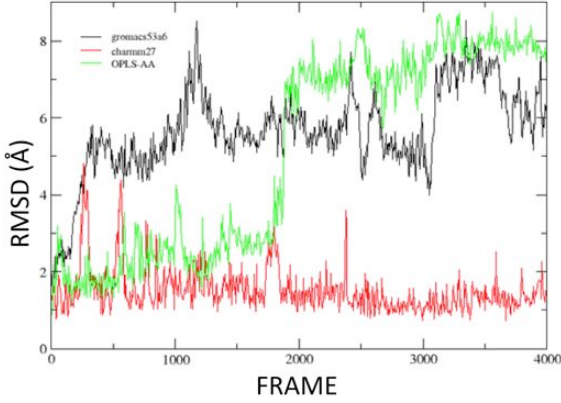


FIG. 7: RMSD of the backbone atoms for the three force fields: GROMOS53a6 (black), OPLS/AA (green), and CHARMM27 (red).

most intense in the GROMOS53a6 force field. With this force field, the initial α -helical structure begins collapsing within 4–5 ns. With the OPLS/AA force field, we find that the deformation starts after around 40 ns. In the case of the CHARMM27 force field, the helix tends to remain intact within the selected time range. The deformation begins only after a substantially longer simulation. Apparently, this force field has a tendency to produce structures that have an overly α -helical content. After extended comparisons of the three force fields, including different time steps and simulation lengths, we have chosen a 80 ns GROMOS53a6 force field trajectory with 2 fs time step, for the final production simulation that we analyse here. Qualitatively, the results that we present are independent of the force field and time step that we have chosen.

Qualitative considerations

In figure 8 we show the results from a `do.dssp` [9] secondary structure analysis, in the case of the GROMOS53a6 simulation. The following *qualitative* observations can be made:

- After around 4–5 ns corresponding to frames 200–250, there is an initial formation of a coil structure, according to `do.dssp` classification. The coil becomes initially stabilised around residue number 29 which corresponds to amino acid number 656 in the PDB file. The coil is connected to the C-terminal with a bend. At around 8 ns (around frame 400) there are helical fluctuations in this structure towards N-terminal, and at around 22–24 ns (frames 1100–1200) the coil structure moves back towards the C-terminal. The motion takes place in two steps, at around frame 1200 and then again at around frame 2800 after which the coil disappears, by merging into the apparently random fluctuations of the C-terminal residues.

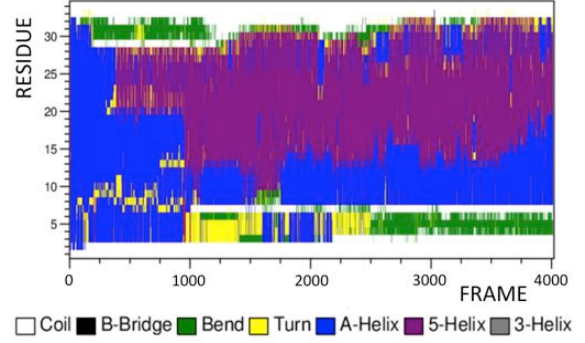


FIG. 8: Secondary structure analysis using `do.dssp` along a 80 ns trajectory, produced using the GROMOS53a6 force field. The PDB residues 628–661 are labeled 0–33.

At the level of `do.dssp` secondary structure analysis the coil which emerges near the C-terminal and propagates along the backbone, is putatively akin the propagating loop structure that has been previously identified and studied in coarse grained UNRES simulations of the protein G related albumin-binding domain with PDB code 1GAB [12, 13]. In particular the UNRES simulation [12] identifies a displaced protein loop as a localised structure with a profile that can be described by the soliton solution of the discrete nonlinear Schrödinger equation (34), (35). The simulation demonstrated that when the loop-soliton moves along the protein lattice, with cells matching the residues, there are waves that are emitted in its wake as vibrations in the lattice structure. These waves drain the kinetic energy of the soliton, and cause it to decelerate. Eventually the kinetic energy of the soliton becomes depleted, and it can no longer cross over the energy barriers between lattice cells and becomes localised around a particular set of lattice cells. The energy barriers that prevent the soliton from translating along the backbone lattice were identified as Peierls-Nabarro barriers [35–37] in [13].

In the present case of the C-terminal coil, there is apparently a Peierls-Nabarro barrier that stops and prevents the coil that is supposedly modelled by a DNLS, from propagating away from the C-terminal beyond the residues 28–29. Instead it becomes initially trapped, then moves towards the C-terminal and dissolves there. The soliton moves step-wise, it’s crossing-over the ensuing Peierls-Nabarro barriers is boosted by thermal fluctuations. The soliton crosses a barrier whenever the amplitude of its thermal fluctuations exceeds the barrier-specific threshold value.

- At around the same time when the C-terminal coil forms, we observe a turn deformation that forms and proceeds away from the N-terminal, and then fluctuates thermally between residues 5 and 10. After around 20 ns (frame 1000) of simulation time, there is a rapid extended turn-like fluctuation that connects the N-terminal with a

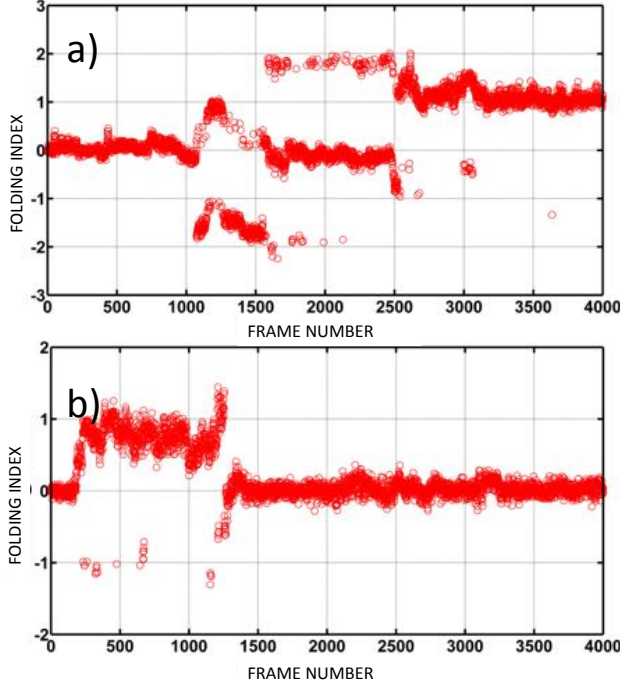


FIG. 9: The folding index density (20) evaluated over two segments of the 1AIK backbone: (a) the segment 4-11 (residues 632-639 in PDB), and (b) the segment 24-30 (residues 652-658 in PDB).

localised structure which is identified as a short coil by `do.dssp`. This is an apparent DNLS soliton, emerging at the end of the extended turn-like structure and stabilising around the residues 6-8 (residues 634-636 in the PDB file). We observe initially relatively strong fluctuations in the residues between the putative soliton and the N-terminal. But the amplitudes of these fluctuations become damped and after around 50 ns (frame 2500) there are only minor fluctuations in the soliton. There is a bend between the soliton and the N-terminus which constitutes a Peierls-Nabarro barrier, high enough to prevent the soliton from moving towards the N-terminal, step-wise by thermal fluctuations.

Backbone folding index

We proceed to analyse the dynamics *quantitatively*, and we start with the backbone folding index (19). For this we have divided the backbone into segments of varying length, and computed the folding index over the segments during the 80 ns time evolution. Examples of results are shown in Figure 9, where we plot the numerical values of the folding index density (20). The first segment consists of the sites 4-11 corresponding to residues 632-639 in PDB. This segment covers the N-terminal soliton structure (see Figure 8). The second segment consists of

the sites 24-30 (652-658 in PDB). This covers the segment where the C-terminal coil initially appears in Figure 8. We observe the following:

Figure 9(a): Initially, the folding index of the segment 4-11 vanishes. But in the vicinity of frame 1000, coinciding with the formation of the N-terminal coil/soliton in figure 8, the folding index starts fluctuating between the values $Ind_f = \pm 1$ and $Ind_f = -2$. We note how the pattern of the oscillations reflects the structural changes in the region between the coil and the N-terminal, shown in figure 8: There are first fluctuations between turn and bend, during frames 1000-1500. When the oscillations in the values of the folding index diminish and vanish near the frame 1500, we observe a formation of helical structure between the coil and the N-terminal in Figure 8. The folding index then starts oscillating between the values $Ind_f = 0, 2$, and this corresponds to a frame segment where the helix converts into a turn in the Figure 8. At around the frame 2500, the folding index finally stabilises to the final value $Ind_f = +1$. This stabilisation concurs with the formation of a bend between the coil and the N-terminal, in Figure 8.

Figure 9(b): We observe the increase of folding index from $Ind_f = 0$ to $Ind_f = +1$ near frame 200, and subsequent decrease back to $Ind_f = 0$ near frame 1300. According to figure 8, these transitions coincide with the appearance of the C-terminal soliton, and its subsequent propagation towards the C-terminal, away from the segment 24-30.

Accordingly, we have found that the variations in the values of the folding index, in particular in the case of the N-terminal soliton, coincide with the structural deformations that take place along the backbone segment which is located between the soliton and the N-terminal. In particular, the final stabilisation of the folding index concurs with the crossing over the Peierls-Nabarro barrier and subsequent stabilisation of the soliton, according to figure 8. Moreover, the C-terminal soliton emerges with $Ind_f = +1$ and remains stable until the Peierls-Nabarro barrier crossing takes place. The evolution of the ensuing folding index is also fully in line with the results that we deduce from `do.dssp`. We conclude that the behaviour in the backbone segment, surrounding the soliton, directly correlates with the topological character of the soliton, in both cases.

Finally, in Figure 10 we show the folding index density (20) over the *entire* backbone and for all the 4000 frames. We observe that:

- Initially, the folding index vanishes. This is consistent with the α -helical structure of the 1AIK sub-chain. But there is a sudden initial transition to the value $Ind_f = +1$, presumably reflecting the initial stages of C-terminal soliton formation.
- Up until the frame ~ 2700 the folding index tends to vanish. But there are fluctuations, mainly between values ± 2 which reflect the various processes that take

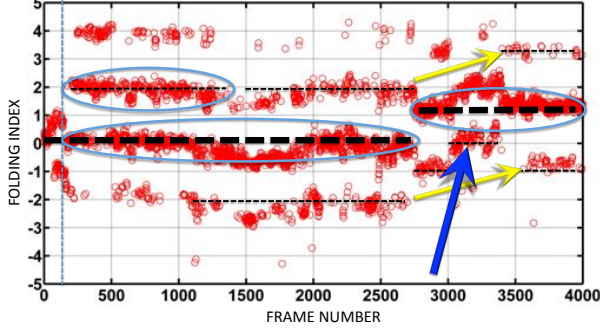


FIG. 10: The folding index density (20) evaluated over the entire backbone for all 4000 frames. Some of the major features have been high-lighted.

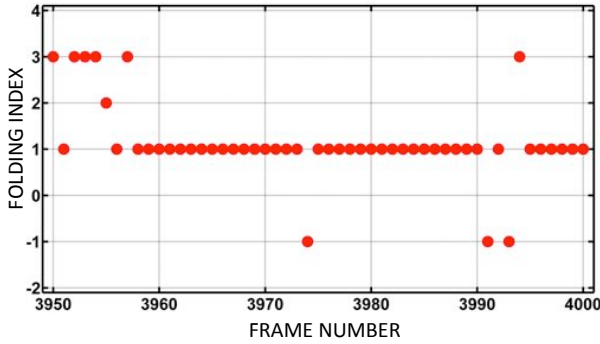


FIG. 11: The close-up of segments 3050-4000 in the folding index density of figure 10.

place near the terminals.

- In the vicinity of frame 2700 there is a transition, and the value of the folding index starts stabilising toward the value $Ind_f = +1$. This stabilisation concurs with the stabilisation of the N-terminal soliton, and the final departure of the C-terminal soliton. The fluctuations also shift, oscillating between $Ind_f = +3$ and $Ind_f = +1$ and this shift is identified by the yellow arrows in the figure.

In Figure 11 we show a close-up to the last 50 frames in figure 10. It confirms the stabilisation of the folding index towards the value $Ind_f = +1$, with occasional fluctuations where $Ind_f = +3$ or $Ind_f = -1$. In Figure 12, following Figure 3(a), we show the full trajectory for the entire final frame 4000. The trajectory starts from the N-terminal which is located in the β -stranded region of Figure 2. It moves over to the α -helical region, then return to the β -stranded region to encircle the north pole. Finally, the trajectory merges and ends with the α -helical region. The trajectory confirms that the final structure at frame 4000 indeed does support a twisting $\Delta\tau = +\pi$ and that the twisting is furthermore located at the N-terminal soliton.

The stabilisation of the folding index to the value $Ind_f = +1$ confirms the global character of the remaining

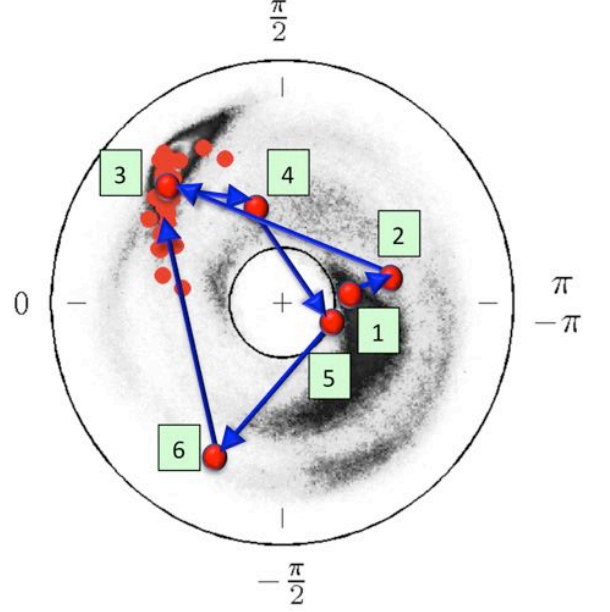


FIG. 12: The trajectory of the frame 4000 on the (κ, τ) landscape of figure 2, following figure 3 a). Note that after residue 6, the remaining residues are all located closely, in the α -helical region.

N-terminal soliton structure: There is a total twisting by $\Delta\tau = +\pi$ along the final backbone, in comparison to the initial configuration and including the terminal residues, and this twisting is localised on the N-terminal soliton. Moreover, we observe that the N-terminal residues are in a β -stranded position while the C-terminal residues are in the α -helical position.

Side chain analysis

Figure 4 shows the landscape of the ground state (crystallographic structure) $C\beta$ atom directions in the $C\alpha$ centered discrete Frenet frames. Figure 13 shows how the directions of the $C\beta$ evolve during our entire GRO-MOS53a6 simulation.

We find it remarkable how similar the dynamical landscape of Figure 13 is with the static ground state landscape shown in Figure 4: The direction of $C\beta$ nutates *tightly* around its static ground state landscape. Clearly, there must be strong correlations between the backbone $C\alpha$ and side chain $C\beta$, during the entire dynamics. Accordingly, the information content in the angles η_i in (16), (17) should correlate strongly with the $C\alpha$ geometry changes during the dynamical process.

More generally, we expect that the various backbone and side chain spin models that we have introduced, are all in the same dynamical universality class.

Figure 14 shows the residue-wise accumulated distribution of all the individual angles η_i in (16) *i.e.* the

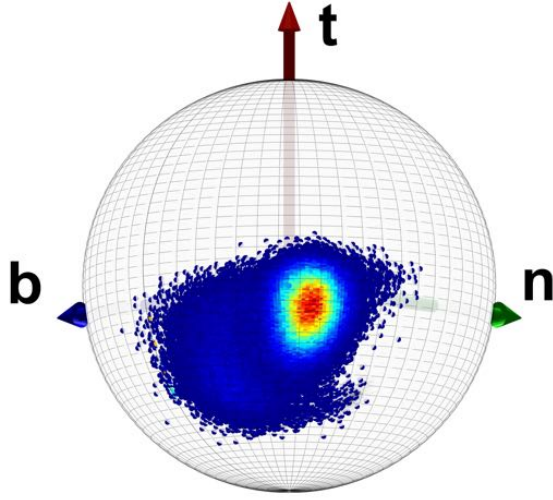


FIG. 13: The dynamical landscape of all the $C\beta$ atoms during the entire 80 ns GROMOS53a6 simulation. A comparison with Figure 4 establishes the presence of strong correlations between the backbone $C\alpha$ and the side chain $C\beta$ geometries during the entire process.

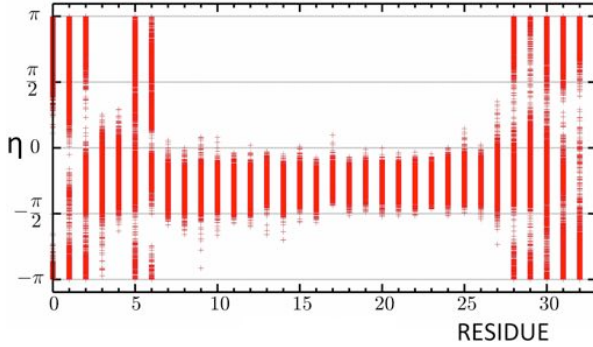


FIG. 14: The residue-wise distribution of the angles η_i in (16), during the entire GROMOS53a6 80 ns run. The horizontal axis labels the residues and the vertical axis is the value of the angle η in radians.

ensuing landscape of the individual η_i during the entire 80 ns GROMOS53a6 simulation. The conclusions that can be deduced from this figure are in line with those in Figure 8. In particular, we observe the presence of the N-terminal soliton, how it is centered around residues 5-6. We also observe that the residues between sites 6-27 are in a helical position during the entire time evolution. We also observe the merging of the C-terminal soliton with the fluctuations of the C-terminal.

Figure 15 shows the time resolved landscape of all the η_i angles. There is a remarkable similarity between this figure, and the figure obtained from the `do.dssp` backbone analysis shown in figure 8. In particular, the formation and stabilisation of the N-terminal soliton around sites 5-6 is clearly visible. The appearance of the C-terminal soliton and its subsequent evolution is simi-

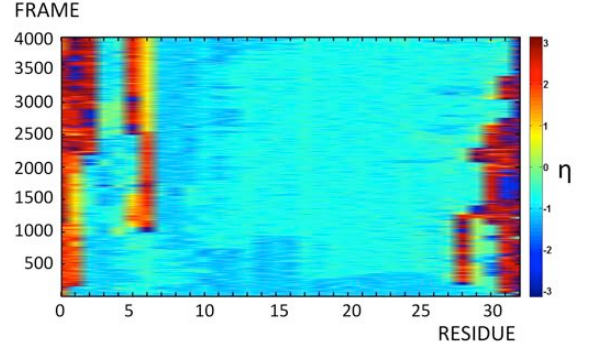


FIG. 15: The time resolved evolution of all the individual angles η_i . Note the similarity with Figure 8.

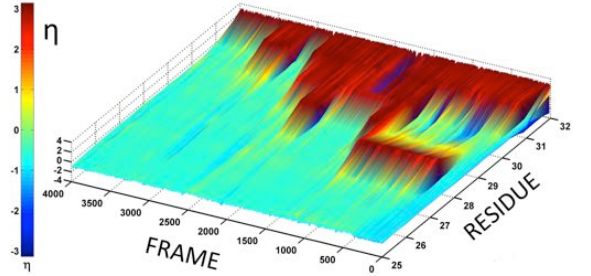


FIG. 16: A close-up of the time resolved evolution of the individual angles η_i , in the case of the C-terminal soliton structure; only the last 9 residues along the backbone are shown.

larly visible: we observe how this soliton is formed at around frame 200, and then propagates towards the C-terminal in a stepwise manner, crossing over the various Peierls-Nabarro barriers and eventually merging with the C-terminal thermal fluctuations.

Finally we note the apparent similarities between the structure of the landscape in Figure 15 and the behaviour of the folding index density in Figure 10.

Details

We proceed to analyse the detailed properties of soliton structure and formation. Our main focus will be on the N-terminal soliton structure. We are particularly interested in the phenomena that take place when the soliton is formed, *i.e.* the vicinity of the frame ~ 1000 , and when the soliton moves over to a Peierls-Nabarro barrier and stabilises, *i.e.* the vicinity of the frame ~ 2500 .

C-terminal side chain soliton

We start with a closeup of the C-terminal part in Figure 15, shown in Figure 16. We observe how, in terms of the side chain η_i angles, the soliton structure which forms

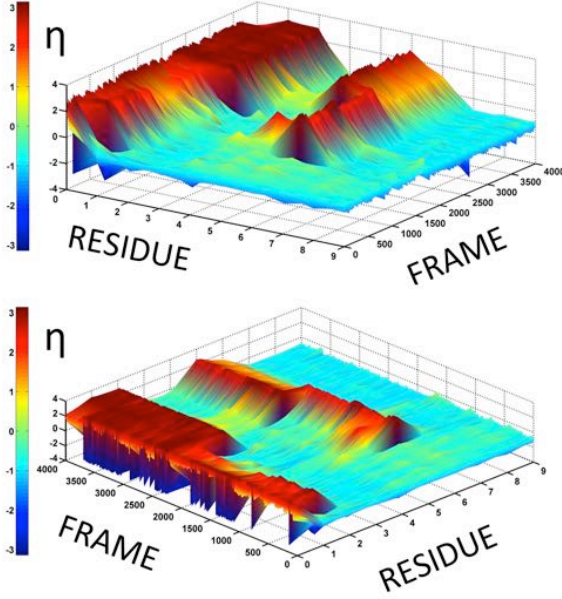


FIG. 17: The time resolved evolution of the individual angles η_i , for the N-terminal soliton structure. The two figures show the same data, but from different perspectives, for the first 10 residues.

with center at residue 28 subsequently propagates back-and-forth, in a step-wise manner, towards the C-terminal. Eventually it merges with the terminal, and dissolves into its fluctuations. The soliton motion is fully in line with Figure 8, and the initial motion is consistent with the folding index analysis in Figure 9(b). In particular, the step-wise propagation of the soliton is fully consistent and in line with the presence of Peierls-Nabarro barriers. These barriers are high enough to trap the soliton momentarily, but low enough for the soliton to eventually cross over them when its thermal excitation energy fluctuates to high enough value.

We remind that the present simulations have been performed at 290 K. We have chosen this relatively low temperature value, from the *in vivo* perspective, in order to restrain the soliton mobility and to dampen noisy thermal fluctuations.

N-terminal side chain soliton

Figure 17 shows a close-up of the N-terminal part in Figure 15, from two complementary perspectives. The soliton appears in the vicinity of frame 1000. It subsequently translates one residue towards the N-terminal, in the vicinity of frame 2500. This is an apparent crossing over a Peierls-Nabarro barrier by thermal fluctuation, and it is followed by a stabilisation of the soliton at the final position. Note the correlation between the soliton motion and the extent of the N-terminal fluctuations.

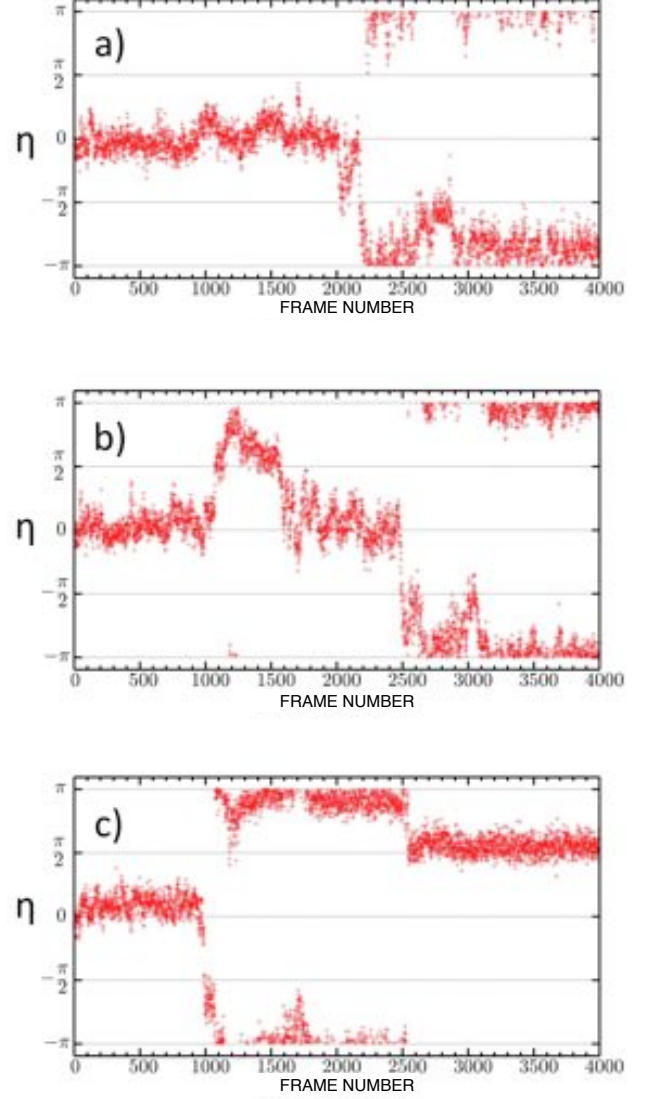


FIG. 18: Fluctuations in the values of the η_i angle during the entire dynamical process, around the average initial value evaluated from the original PDB structure. Panels (a)–(c) correspond to residues $i = 2, 5, 6$ respectively. Note that the values are in the range $[-\pi, \pi] \bmod(2\pi)$.

Figure 18 shows the evolution of the individual η_i angles in (21), in the case of the N-terminal soliton structure. The panels display the deviation of η_i from the initial average value for residues $i = 2, 5, 6$, which we have found to be those of primary interest. For $i = 3, 4$ and for $i = 7$ and larger, the deviations from the initial average value fluctuate around zero.

In Figure 19 we scrutinise those segments of Figure 18, where the $\bmod(2\pi)$ branch of the angle needs to be carefully resolved. In each of the panels in Figure 19 we display the data in Figure 18, over a subset of frames and now evaluated around the value $\pi/2$: in Figure 19(a) we show the segment 1500-2500 of Figure 18(a), in Figure

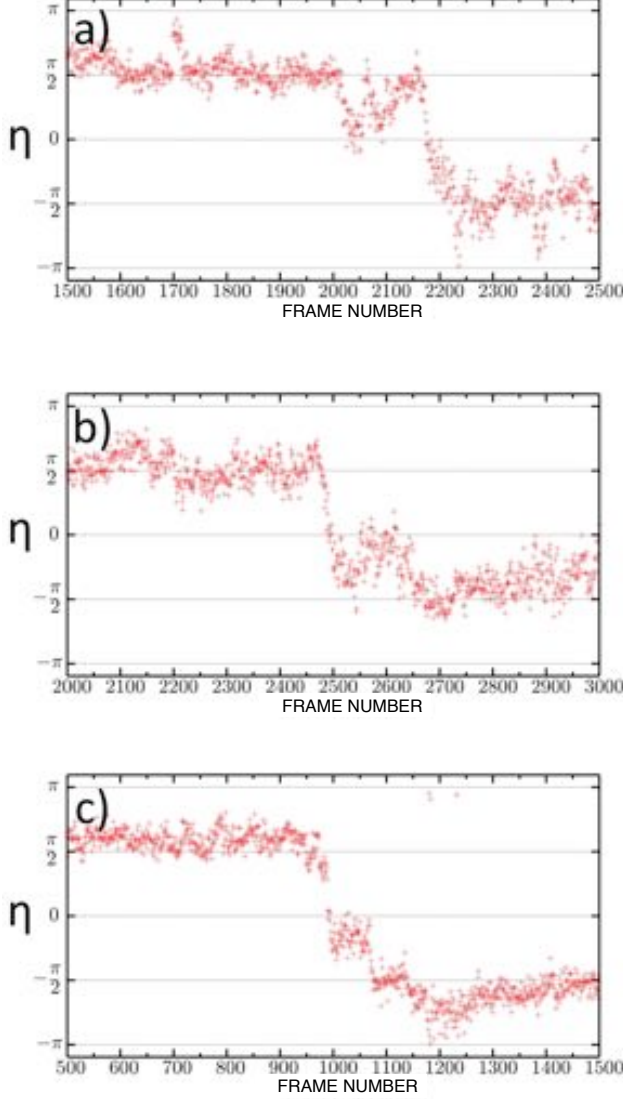


FIG. 19: Details of the corresponding panels in Figure 18, where a careful scrutiny is needed due to the multivaluedness of the angle. For this, we display the fluctuations of the ensuing angle around the value $\pi/2$ as follows: in panel (a) we have details of Figure 18(a) over the frames 1500-2500, in panel (b) we have details of Figure 18(b) over the frames 2000-3000, and in panel (c) we have details of Figure 18(c) over the frames 500-1500.

19(b) we show the segment 2000-3000 of Figure 18(b) and in Figure 19(c) we show the segment 500-1500 of Figure 18(c).

By combining Figures 18 and 19 we conclude that there are two major transitions, around frames 1000 and 2500 respectively. These transitions are concurrent with the major transitions in figures 8, 15, 17 and in particular figure 9. The first transition corresponds to the creation of the N-terminal soliton, and the second one to its translocation, by one site towards the N-terminal, and subse-

quent stabilisation. We also observe the presence of an extended transition process, visible in Figure 18(b) between frames 1100-1500. In summary, we conclude from these figures that:

- In the vicinity of the frame 1000, when the N-terminal soliton structure forms, there is an initial twisting of the $i = 5$ dihedral which is close to $+\pi$ and a twisting of the $i = 6$ dihedral by an approximatively equal amount but in the opposite direction. Thus, at this point the total twisting which is produced along the side chain segment vanishes. We note that the presence of two twists by an equal amount but opposite in direction, is the hallmark of a Bloch domain wall pair production. But we recall that the backbone folding index detects only a single soliton, as shown in Figure 9(a).
- After the initial Bloch wall pair formation, the $i = 5$ dihedral becomes slowly twisted back to the original value between frames 1200-1600, so that after frame 1600 we are left with a total of $\sim -\pi$ twist. This process leaves us with a single soliton along the chain segment, with a total twisting around $-\pi$.
- Finally, in the vicinity of frames 2100-2200, we observe a rapid twisting of the $i = 2$ dihedral by an amount close to $\approx -\pi$. This is followed, in the vicinity of the frame 2400-2500, by a twisting of the $i = 5$ dihedral by an approximatively equal amount. There is an accompanying twist of the $i = 6$ dihedral by an amount somewhat less than $-\pi/2$, over the same frames. These two twistings at $i = 5, 6$ accompany the Peierls-Nabarro barrier crossing, as can be deduced by comparison with figures 15 and 17.
- When the soliton stabilises, after frame 3000, we conclude that there is a total twisting in the side chain structure which is close to $\approx -3\pi$, and carried by the final soliton configuration. Thus we assign to the final soliton the value $Ind_m = -3$ of the index (22).

Note that there is also certain (small) spillage of the η -values which is distributed among the nearby residues.

In summary, the side chain analysis shows that at the level of the XY spin chain analysis, the N-terminal soliton structure forms by an initial rapid formation of a Bloch domain wall *i.e.* soliton-antisoliton pair, followed by a slow twisting that apparently removes one of the two Bloch wall solitons. This is then followed by a rapid transition, in combination with a Peierls-Nabarro barrier crossing, that forms the final stable soliton structure. Accordingly, we may characterise the final soliton as a configuration with the $C\alpha$ backbone folding index $Ind_f = -1$ and the side chain XY folding index $Ind_m = -3$.

Backbone

Figure 13 reveals the presence of strong correlations between backbone and side chain dynamics. In particu-

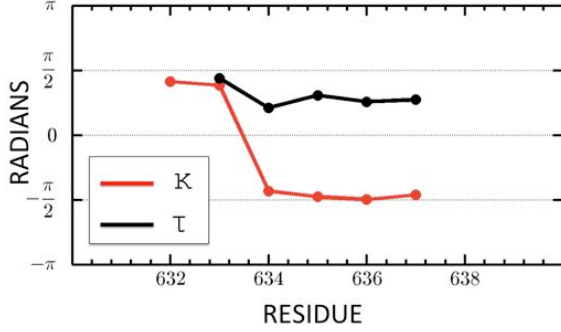


FIG. 20: The \mathbb{Z}_2 gauge-transformed bond angles κ_i (red) and torsion angles (black) for the segment 4-11 (PDB index 632-639) in frame 2000. Note that bond angle takes 3 residues, and torsion angle takes 4 residues to compute.

lar, any formation of side chain soliton structure should correlate with the formation of corresponding backbone soliton. Accordingly, we proceed to construct explicitly the backbone soliton that accompanies the N-terminal side chain soliton. We shall find that the backbone soliton can be modeled by a solution of the discrete nonlinear Schrödinger equation (34), (35), with very high sub-atomic precision.

As an example, we consider the profile of the N-terminal soliton structure at two different frames. Other frames, and the C-terminal soliton structure, can be analysed similarly. We select the sites 4-11 (PDB sites 632-639) for our analysis: The sites 0-3 are subject to fluctuations, and beyond the site 11 there is only a monotonic α -helix.

We start with the frame 2000, which is located in the regime where the side chain structure of the soliton has stabilised according to Figure 18(b) and the backbone folding index of the segment shown in Figure 9(a) has the value $Ind_f = +1$; the side chain index over this segment is $Ind_m = -1$, according to Figures 18, 19.

In Figure 20 we show the profile of the bond angle and the torsion angle over the segment 4-11 (sites 632-639 in the PDB file) in the frame 2000, after we have implemented the \mathbb{Z}_2 gauge transformation (11) to identify the soliton profile. Note that in order to compute a single bond angle, we need to know three residues while the evaluation of a torsion angle consumes four residues. Thus, despite the smaller number of data points in the figure, the ensuing configuration engages 8 residues.

We observe that the bond angle has the profile of a single domain wall soliton of the DNLS equation, approximated by (31). We use the software package ProPro that has been described at

<http://www.protein-folding.org>

to numerically construct the ensuing soliton solution of the DNLS equation.

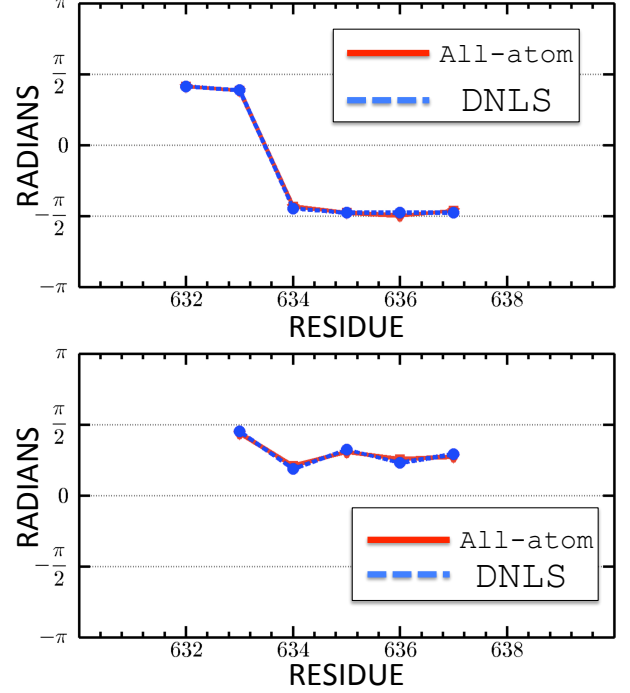


FIG. 21: The \mathbb{Z}_2 gauge-transformed all-atom bond angles κ_i (top) and torsion angles (bottom) for the segment 632-639 in frame 2000, compared with the corresponding DNLS soliton, Eqs. (34), (35).

In Figure 21 (a) we compare the profile of the bond and torsion angles in Figure 20 with the profile of the soliton solution of (34), (35).

In Figure 22 (top) we compare the residue-wise distance between the all-atom configuration of frame 2000, and the DNLS solution. The average $C\alpha$ RMSD between the two configurations is less than 0.1 Ångström, and at no residue is the distance between the $C\alpha$ -atoms more than 0.2 Ångström: the difference is truly negligible. The grey strip around the DNLS soliton is a 0.2 Ångström (quantum mechanical) fluctuation band [27]. The Figure 22 (bottom) shows the 3D overlay of the ensuing $C\alpha$ backbones, for all practical purposes they are the same.

In Figure 23 we show the loop trajectories of the all-atom configuration of frame 2000, and the corresponding DNLS soliton, on the stereographically projected two-sphere of Figure 2 and 3. Note that for both of these two loop trajectories the folding index, as defined in (19) is vanishing, in that the trajectory does not encircle the north-pole (center of the disk). A very short change either in the position of the residue labeled *B* or in the position of the residue labeled *C* in Figure 23 a), can shift the trajectory so that the line connecting them moves over to the other side of the north pole and the folding index becomes $Ind_f = +2$. This is consistent with the result shown in Figure 9 that the folding index fluctuates between the values $Ind_f = 0$ and $Ind_f = +2$, around the

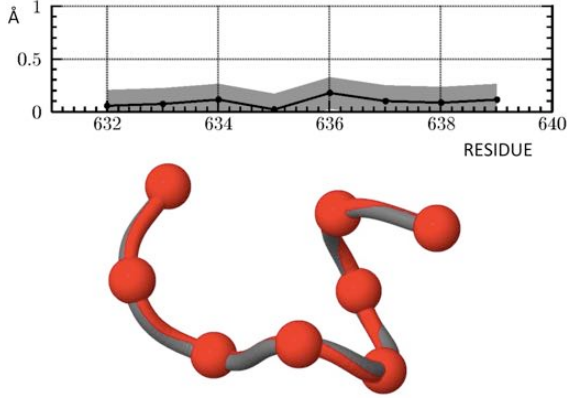


FIG. 22: Top panel: The residue-wise distance between C α backbone of the MD simulated soliton at frame 2000, and the DNSL soliton solution; the grey strip represents the estimated 0.2 Å quantum mechanical fluctuation band; Bottom panel: The 3D superimposition of the all-atom structure (grey) with the DNLS soliton (in red).

frame 2000, with the posture shown in Figure 23 being the more stable one.

In Figure 24 we show a close-up to the frame segment 2475-2525 around the soliton 2000, in terms of the side chain angles η . The close-up reveals the presence of fluctuations between the soliton and the N-terminal, while the helix between the soliton and the C-terminal displays very small fluctuations. Thus, the fluctuations in the folding index around the frame 2000 are most likely due to shifts in the position of the residue labeled B in Figure 23 (a).

Figures 25-28 show the same analyses for the conformation in frame 3500. We find that the DNLS soliton describes the domain wall soliton that we have constructed by all-atom simulations, with a very high sub-atomic precision. We note that the soliton in frame 3500 is a configuration that connects between the β -stranded region of Figure 2 to the α -helical region, while in the case of the soliton at frame 2500 the initial residue is located in a sparsely populated region of the landscape in Figure 2.

The major qualitative difference between frames 2000 and 3500 is between Figures 23 and 28. The soliton in frame 3500 is relatively stable. In particular, as shown in Figure 9, its folding index $Ind_f = +1$. We can understand the stability of the folding index by comparing Figure 23 (a) with Figure 28 (a). In the later, the residues have assumed positions where the connecting arrows are stabilised against small perturbations, they are protected from crossing over the north pole in a manner that causes fluctuations in the value of the folding index.

Finally, we compare the structure of the solitons at frames 2500 and 3500. From Figures 20, 25 we observe that in terms of the bond angles the soliton 3500 has indeed moved one site towards the N-terminal, from the position of soliton 2500. In Figure 29 we overlay their

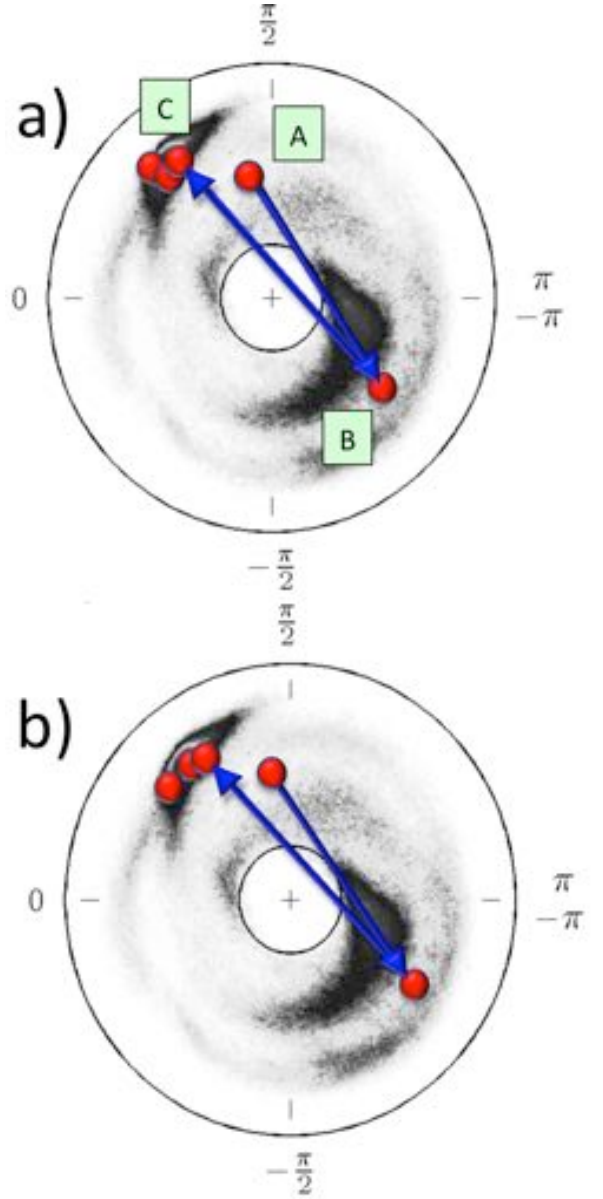


FIG. 23: top: The loop trajectories of Figure 3 (a) in the case of: (a) the all-atom frame 2000 segment 632-639, and (b) the corresponding DNLS soliton.

3D structures. For this, we first translate the soliton in frame 3500 one residue away from the N-terminal, so that the two have the same location along the backbone. The figure shows the ensuing 3D interlaced C α backbones, in a relative position where the RMSD is minimal. There is a visible difference, and the minimal RMSD is 2.0 Å; the soliton has clearly become deformed.

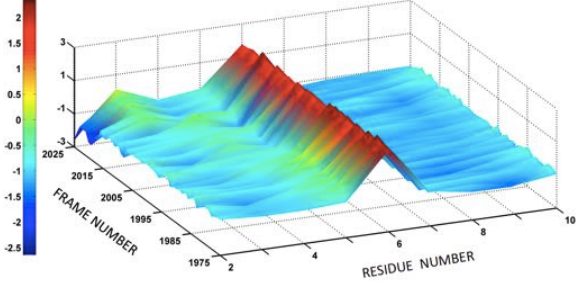


FIG. 24: A close-up of Figure 17, around the frame 2000. The α -helical structure between the soliton and the C-terminal displays only very slight fluctuations, while the fluctuations between the soliton and the N-terminal are more profound.

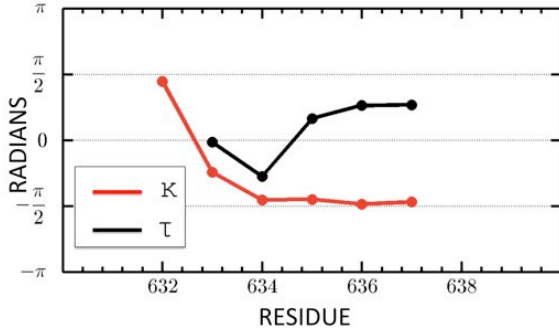


FIG. 25: Same as in Figure 20, for the frame 3500.

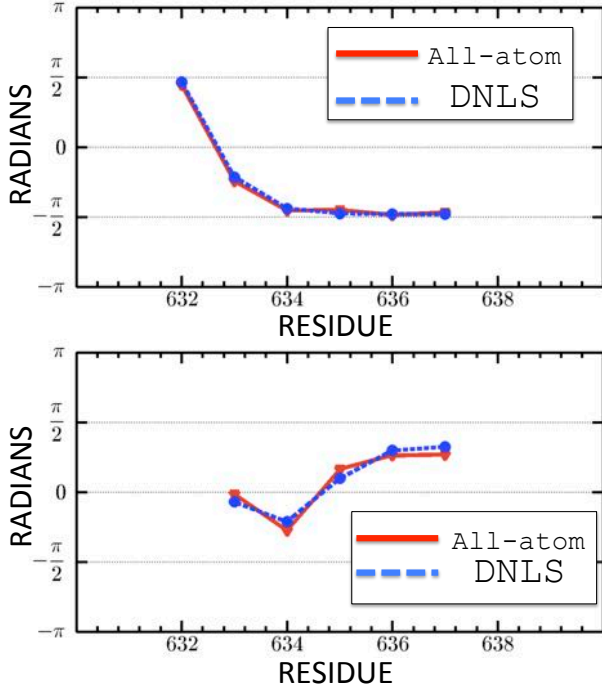


FIG. 26: Same as in Figure 21, for the frame 3500.

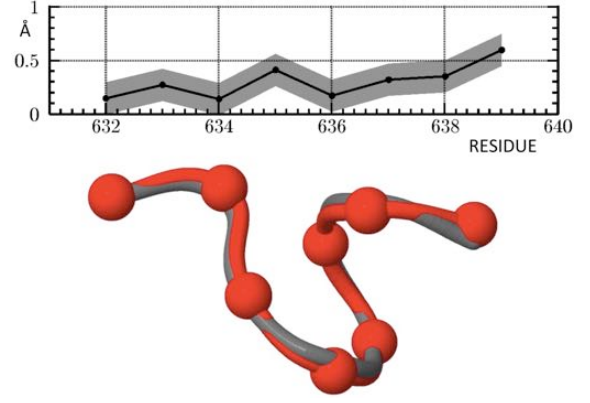


FIG. 27: Same as in Figure 22, for the frame 3500.

SUMMARY

Molecular dynamics enables the scrutiny of protein folding, at the level of individual atoms and over very short time intervals. However, it can leave us with the conceptual challenge to understand, how the individual atoms cooperate to produce the kind of large scale organisation that appears to be prevalent among crystallographic protein structures.

We have performed detailed molecular dynamics simulations, with the aim to find out how organised structure emerges when a protein folds. We have first compared three different force fields using the GROMACS 4.6.3. package, to select the proper tools. We have chosen a C-chain subunit from HIV envelope glycoprotein with PDB code 1AIK as a concrete example, partly due to its biomedical relevance even though this is an issue which has not been addressed by us. We have introduced and further developed various tools of modern theoretical physics, to systematise and analyse the data. These include topological tools, conceptual analogies drawn from the notion of spin chains, the notion of Wilsonian universality, and methods based on the analytical structure of the discrete nonlinear Schrödinger equation. In this manner we have arrived at the conclusion that the protein folding is a process that relates intimately to the emergence and interactions of solitons. In particular, a configuration such as the Bloch domain wall along a spin chain appears to be most useful in comprehending how structure emerges and self-organises when a protein folds.

We have inspected both the static and dynamic properties of domain wall solitons and observed that concepts which are familiar from the study of lattice systems, such as the Peierls-Nabarro barrier, also appear along protein backbones lattices, and in fact assume a central role in dictating how the folding proceeds. We hope that our observations help to pave a way for the powerful analytical and topological tools and techniques that have been introduced and developed in the context of integrable

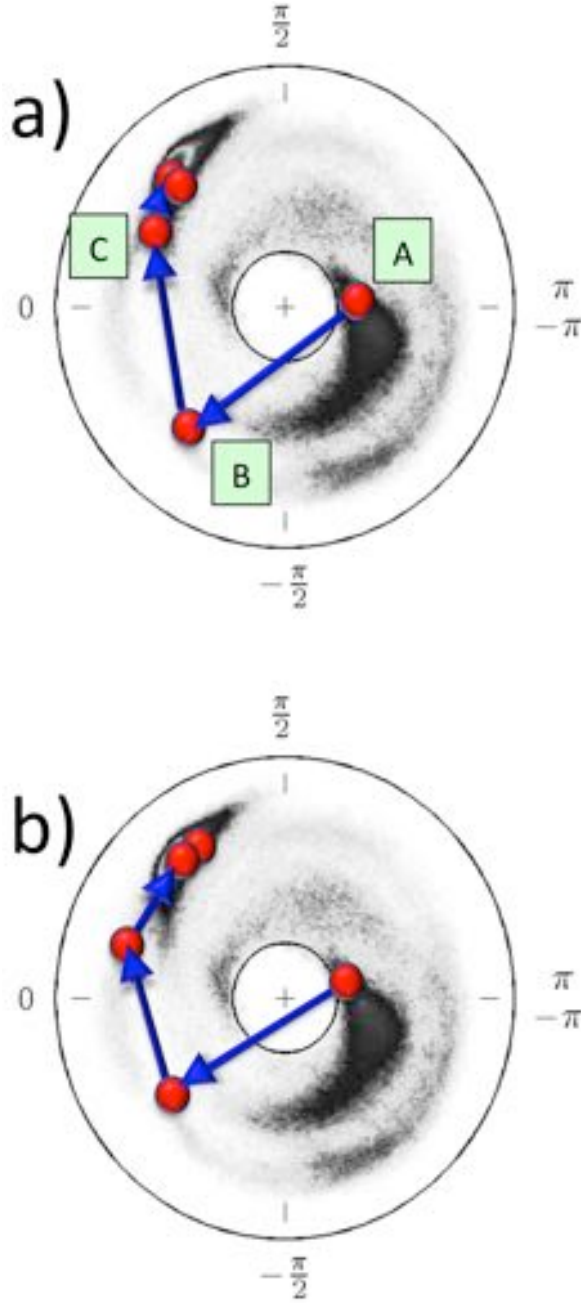


FIG. 28: Same as in Figure 23, for the frame 3500.

spin chains and related solvable models, to become part of the arsenal used describe emergence of structure and organisation in the case of proteins and other biological macromolecules.

ACKNOWLEDGEMENTS:

This research was supported in part by Bulgarian Science Fund (Grant DNTS-CH 01/9/2014) and China-

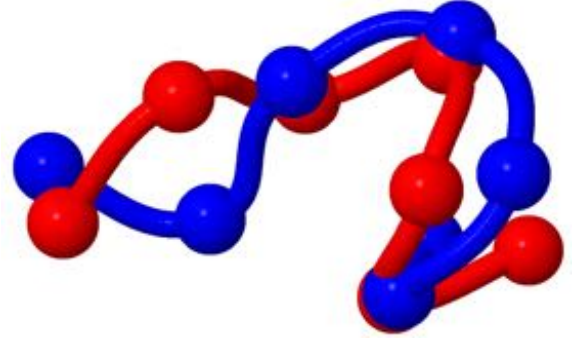


FIG. 29: Comparison of the solitons at frame 2500 (red) and 3500 (blue), after the second has been translated back to the location of the first one. The RMDS is 2.0 Å.

Bulgaria Intergovernmental S&T Cooperation Project at Ministry of Science and Technology of P.R. China (2014-3). AJN acknowledges support from Vetenskapsrådet, Carl Trygger's Stiftelse för vetenskaplig forskning and Qian Ren Grant at BIT, P.R. China. AS was supported by National Science Center Poland (Maestro UMO-2012/06/A/ST4/00376).

* Electronic address: daijing491@gmail.com

† Electronic address: Antti.Niemi@physics.uu.se

‡ Electronic address: hjf@bit.edu.cn

§ Electronic address: adams86@wp.pl

¶ Electronic address: nilieval@mail.cern.ch

- [1] N. Manton and P. Sutcliffe, *Topological Solitons* (Cambridge University Press, Cambridge, 2004)
- [2] L.D. Faddeev, L.A. Takhtajan, *Hamiltonian methods in the theory of solitons* (Springer Verlag, Berlin, 1987)
- [3] M.J. Ablowitz, B. Prinari and A.D. Trubatch, *Discrete and Continuous Nonlinear Schrödinger Systems* (Cambridge University Press, Cambridge, 2004)
- [4] K.A. Dill, J.L. MacCallum Science **338** 1042 (2012)
- [5] A. Leach *Molecular Modeling: Principles and Applications* (Pearson Education, EMA, XX 2001).
- [6] D. Frenkel and B. Smith *Understanding Molecular Simulations* (Academic Press, XX 2001)
- [7] D.C. Chan, D. Fassa, J.M. Berger, P.S. Kim Cell **89** 263 (1997)
- [8] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, P. Bourne, Nucl. Acids Res. **28** 235 (2000); <http://www.pdb.org>
- [9] B. Hess, C. Kutzner, D. van der Spoel, E. Lindahl J. Chem. Theory Comput. **4** 435 (2008)
- [10] N.Molkenthin, S. Hu, A.J. Niemi Phys. Rev. Lett. **106** 078102 (2011)
- [11] A. Krokhotin, A. Liwo, A.J. Niemi, H.A. Scheraga J. Chem. Phys. **137** 035101 (2012)
- [12] A. Krokhotin, A. Liwo, G.G. Maisuradze, A.J. Niemi,

- H.A. Scheraga, J. Chem. Phys. **140** 025101 (2014)
- [13] A. Sieradzan, X. Peng, A.J. Niemi, Phys. Rev. **E90** 062717 (2014)
- [14] M. Khalili, A. Liwo, F. Rakowski, P. Grochowski, H.A. Scheraga, J. Phys. Chem. **B109** 13785 (2005)
- [15] M. Khalili, A. Liwo, A. Jagielska, H. A. Scheraga, J. Phys. Chem. **B109** 13798 (2005)
- [16] A. Liwo, M. Khalili, H.A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **102** 2362 (2005)
- [17] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, J. Hermans, in *Intermolecular Forces* B. Pullman, ed. (Springer, Netherlands, 1981)
- [18] B. Hess, H. Bekker, H.J.C. Berendsen, J.G.E.M. Fraaije, J. Comp. Chem. **18** 1463 (1997) [17] T. Darden, D. York, and L. Pedersen (1993) J. Chem. Phys. **98**, 10089?10092.
- [19] T. Darden, D. York, L. Pedersen J. Chem. Phys. **98** 10089 (1993)
- [20] S.Hu, M. Lundgren, A.J. Niemi Phys. Rev. **E83** 061908 (2011)
- [21] K. Hinsén, S. Hu, G.R. Kneller, A.J. Niemi, J. Chem. Phys. **139** 124115 (2013)
- [22] L.P. Kadanoff, Physics **2** 263 (1966)
- [23] K.G. Wilson, Phys. Rev. **B4** 3174 (1971)
- [24] A.J. Niemi, arXiv:1412.8321 [cond-mat.soft] (2014)
- [25] M. Chernodub, S. Hu, A.J. Niemi, Phys. Rev. **E82** 011916 (2010)
- [26] S. Hu, A. Krokhotin, A.J. Niemi, X. Peng, Phys. Rev. **E83** 041907 (2011)
- [27] A. Krokhotin, A.J. Niemi, X. Peng Phys. Rev. **E85** 031906 (2011)
- [28] M. Lundgren, A.J. Niemi, F. Sha, Phys. Rev. **E85** 061909 (2012)
- [29] M. Lundgren, A.J. Niemi, Phys. Rev. **E86** 021904 (2012)
- [30] X. Peng, A. Chenani, S. Hu, Y. Zhou, A.J. Niemi, BMC Struct. Biol. **14** 482 (2014)
- [31] M. Lundgren, A. Krokhotin, A.J. Niemi, Phys. Rev. **E88** 042709 (2013)
- [32] A.J. Niemi, Phys. Rev. **D67** 106004 (2003)
- [33] U. Danielsson, M. Lundgren, A.J. Niemi Phys. Rev. **E82** 021910 (2010)
- [34] S. Coleman, E. Weinberg, Phys. Rev. **D7** 1888 (1973)
- [35] R. Peierls, Proc. Phys. Soc. **52** 34 (1940)
- [36] F. Nabarro, Proc. Phys. Soc. **59** 256 (1947)
- [37] F. Nabarro, Mat. Sci. Eng. **A234** 67 (1997)